Fermilab

Managed by Fermi Research Alliance, LLC for the U.S. Department of Energy Office of Science

Exascale and Exabytes: Future directions in HEP Software and Computing

Oliver Gutsche

DPF2015 - Meeting of the Division of Particles & Fields of the American Physical Society

6. August 2015

Disclaimer

- About me
  - Scientist at Fermilab
    - Searching for SuperSymmetry and Dark Matter and doing Standard Model Top Physics with CMS
  - Assistant Head of the Scientific Computing Division at Fermilab

- Disclaimer
  - Not a comprehensive review ➜ selection of concepts and developments I think will be important for the future
  - My expertise is in computing for collider experiments, there will be some bias in this talk

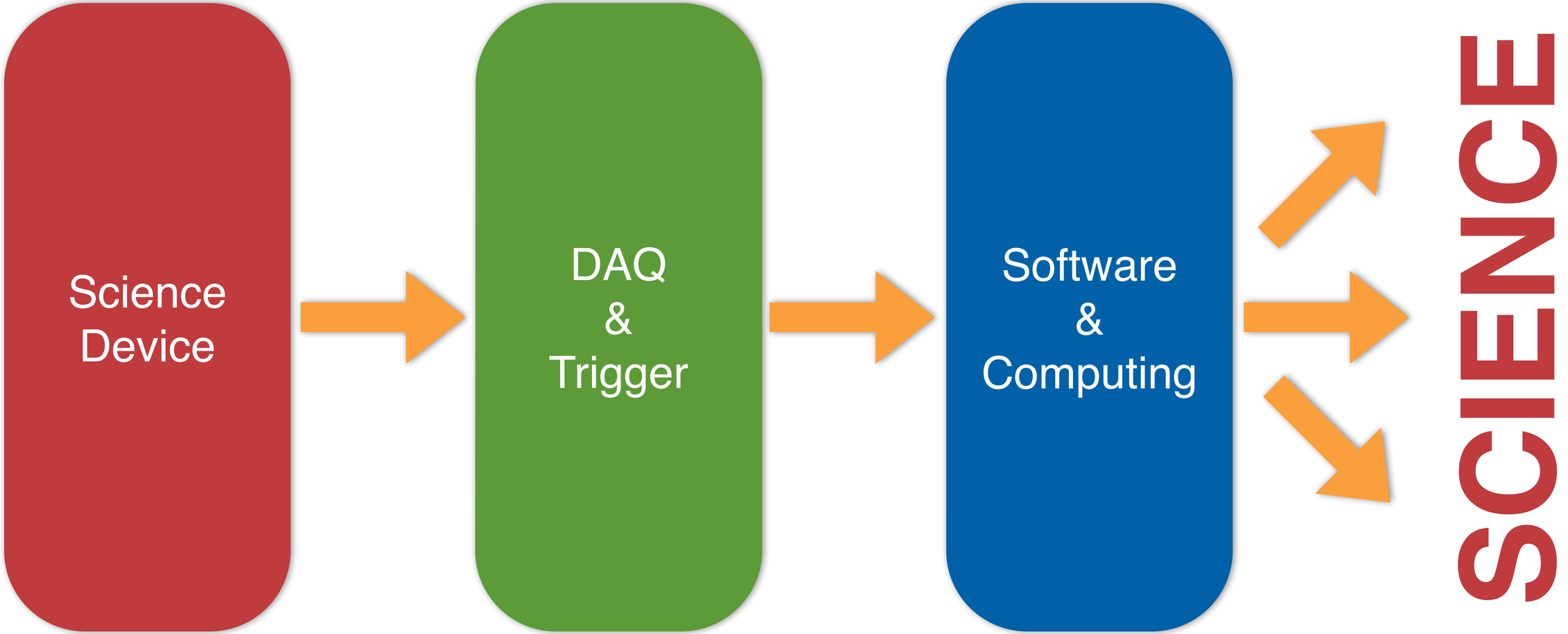
Audience

This talk is for you!

Grad Students

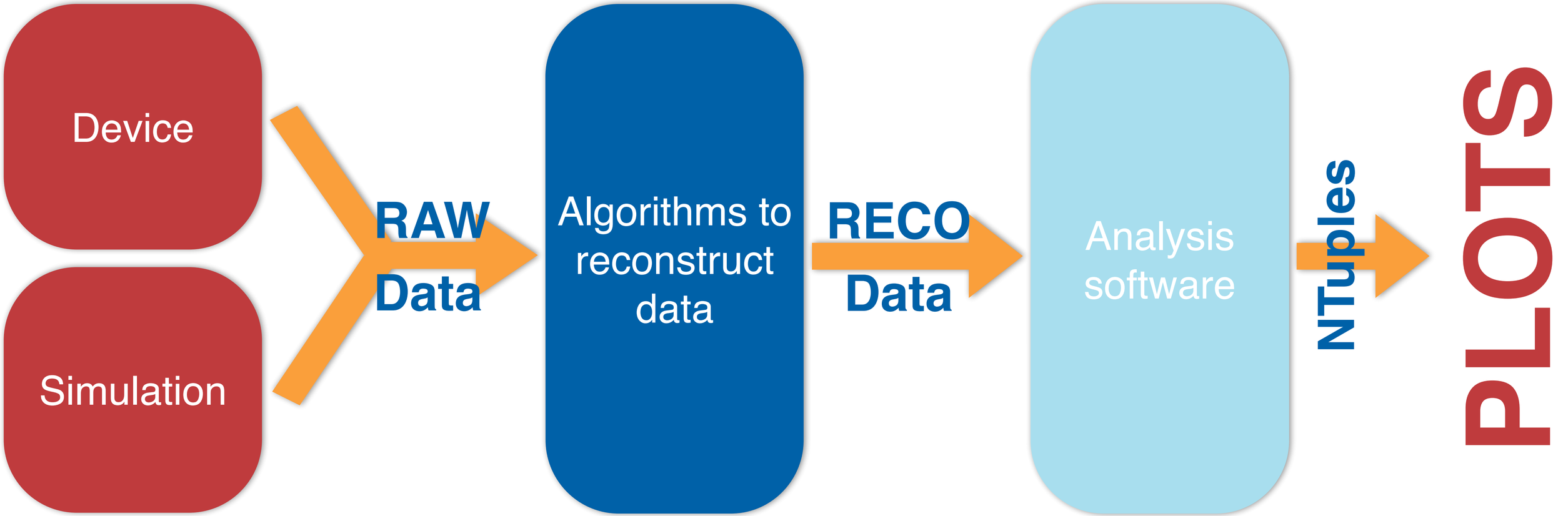
Postdocs

The Scientific Process



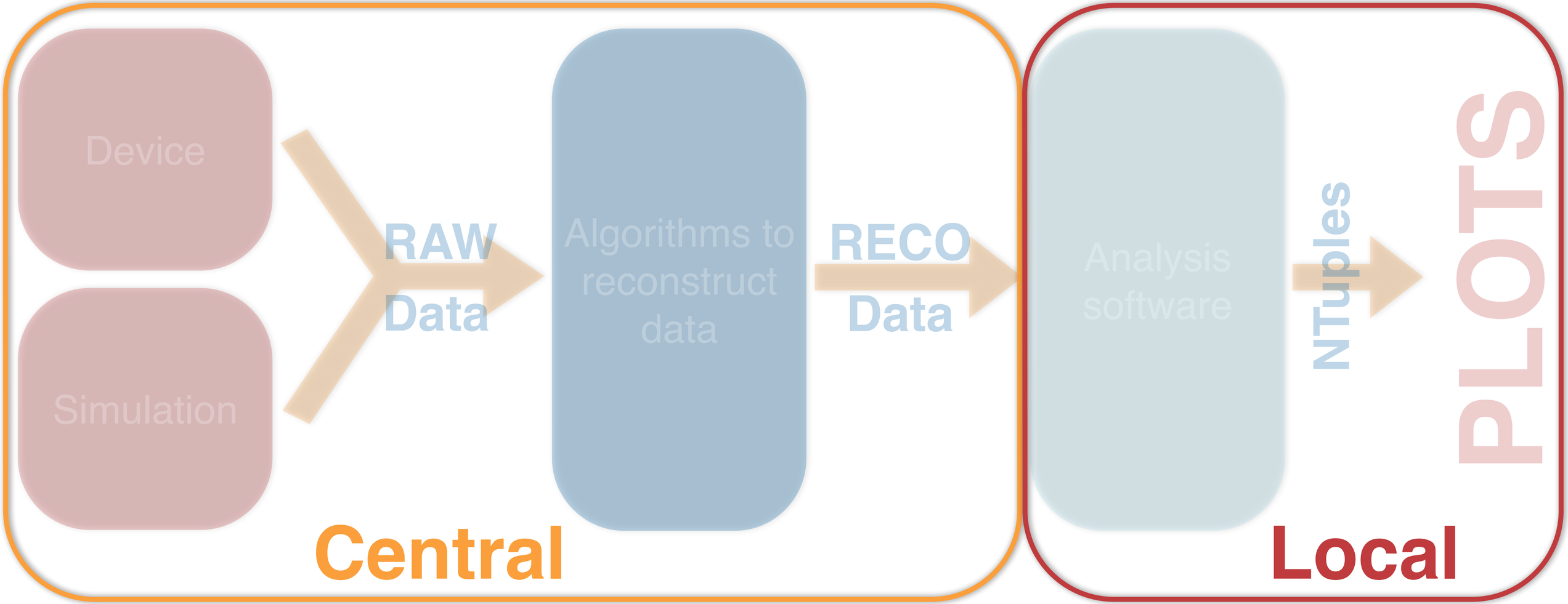
- Software & Computing is an integral part of the scientific process

Software & Computing



- **Software** is important for every step on the way to scientific results

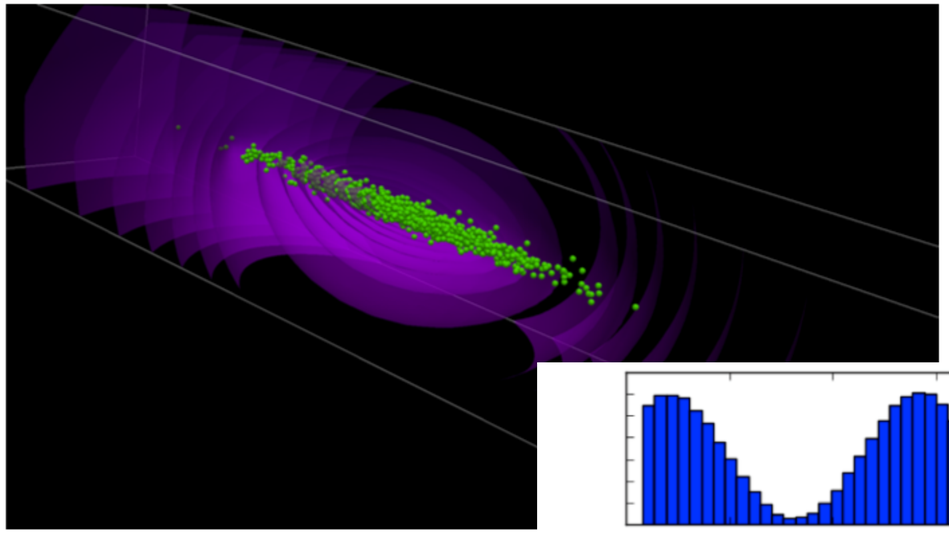
Software & Computing

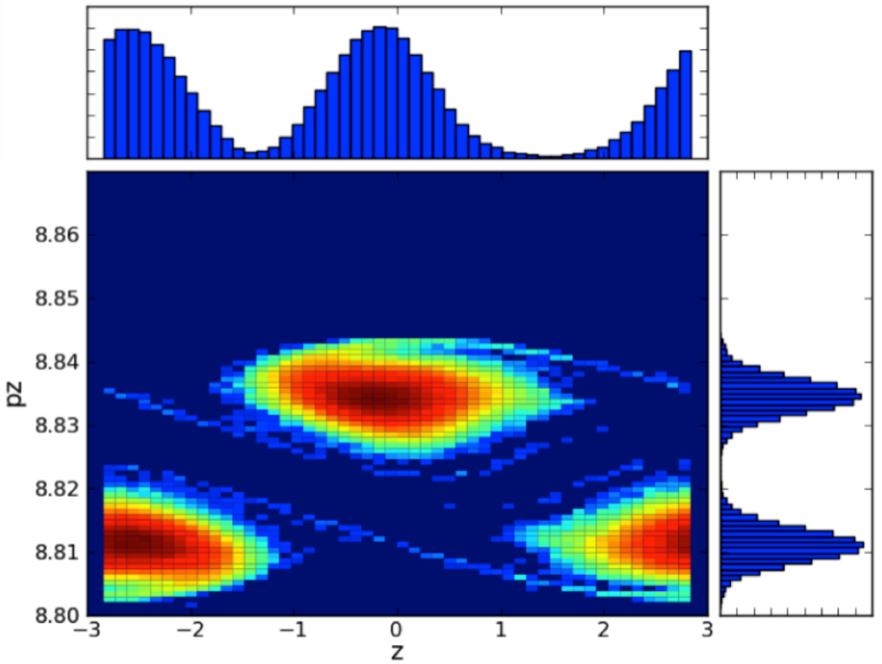


- Computing resources (**Storage and Network**, **Processing**, …) are needed for all steps

Unfair!

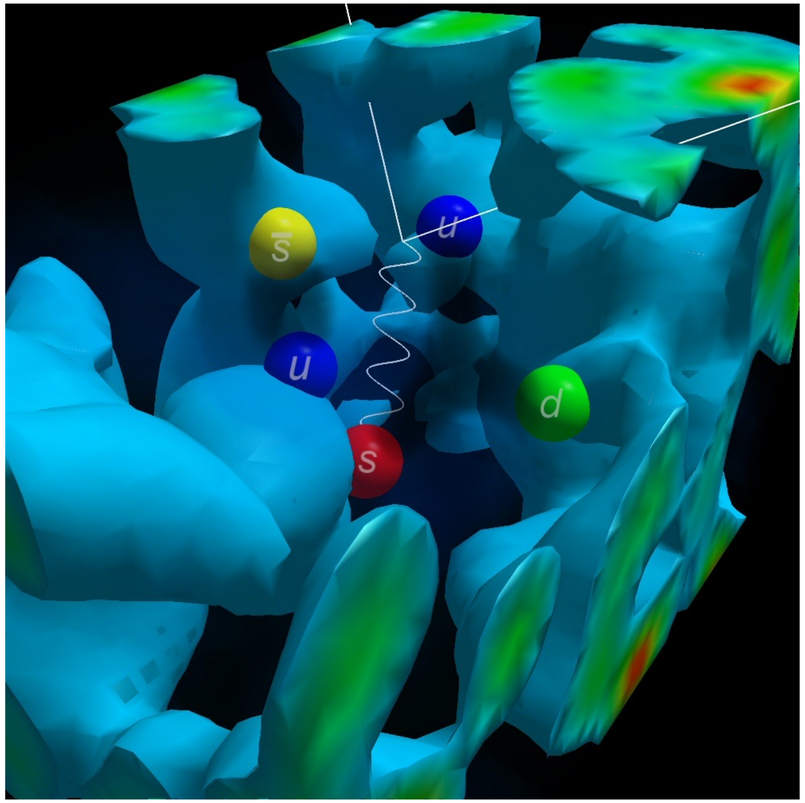
- Simplified picture, I forgot major software & computing areas





Accelerator Simulations

Lattice QCD



everything else I could not include …

Fermilab

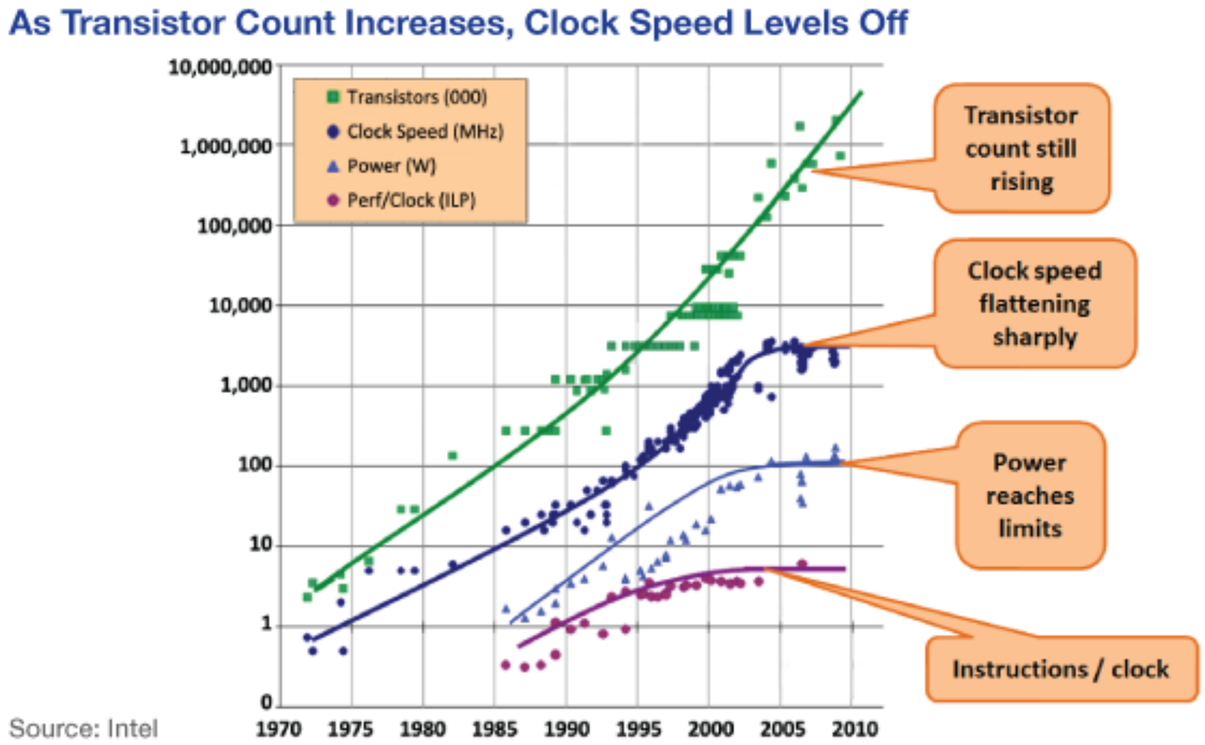
Software

Frameworks

- Underlying infrastructure, core of the software
  - Large experiments have their own Frameworks
  - Trend: community frameworks serving several experiments or detector technologies
    - Art: common framework for neutrino and muon experiments
    - LArSoft: common framework for liquid argon TPC (LArTPC) reconstruction software
    - Gaudi: common underlying framework for ATLAS and LHCb software
    - ALFA: the new ALICE-FAIR software framework
    - …

Moore's Law

- Traditionally, HEP software is optimized for a "simple" architecture
  - x86 based Linux
  - Machines:
    - ≥1 CPUs with ≥1 Cores
  - Shared memory
  - Shared local disk space
  - An application uses one core and memory and local disk space



What we see: more and more cores, but less powerful individually.

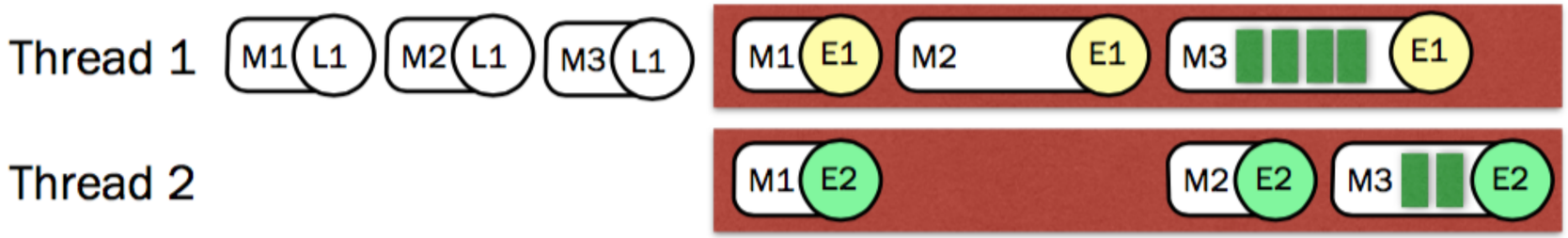
Fermilab

New technologies: more and more cores!

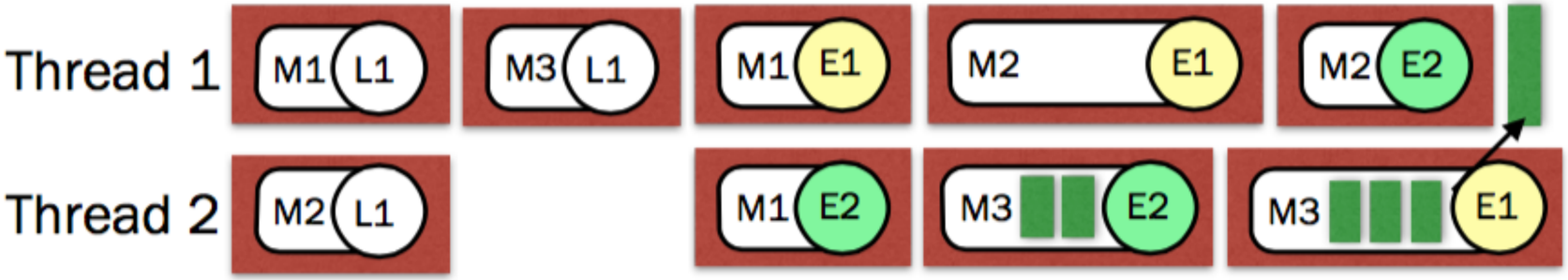
- x86-based machines: running into limitations
  - Each application needs
    - “A lot” of memory (~2GB for LHC experiments) and corresponding bandwidth from memory to a core
    - The more cores in a single machine ➜ the more memory and bandwidth is needed
- New technology: GPGPU: General-purpose computing on graphics processing units
  - Use of a graphics processing units (GPUs) optimized for parallel processing ➜ using many cores per application
  - To perform computation traditionally handled by the central processing unit (CPU)
- New technology: Co-Processor architectures
  - Keyword: Intel MIC (Many Integrated Core) Architecture
- Consequence: **We need to use more cores in parallel for our applications!**

Multi-threading: frameworks

- Advantage: save memory by sharing between threads
- current state: run each event in own thread



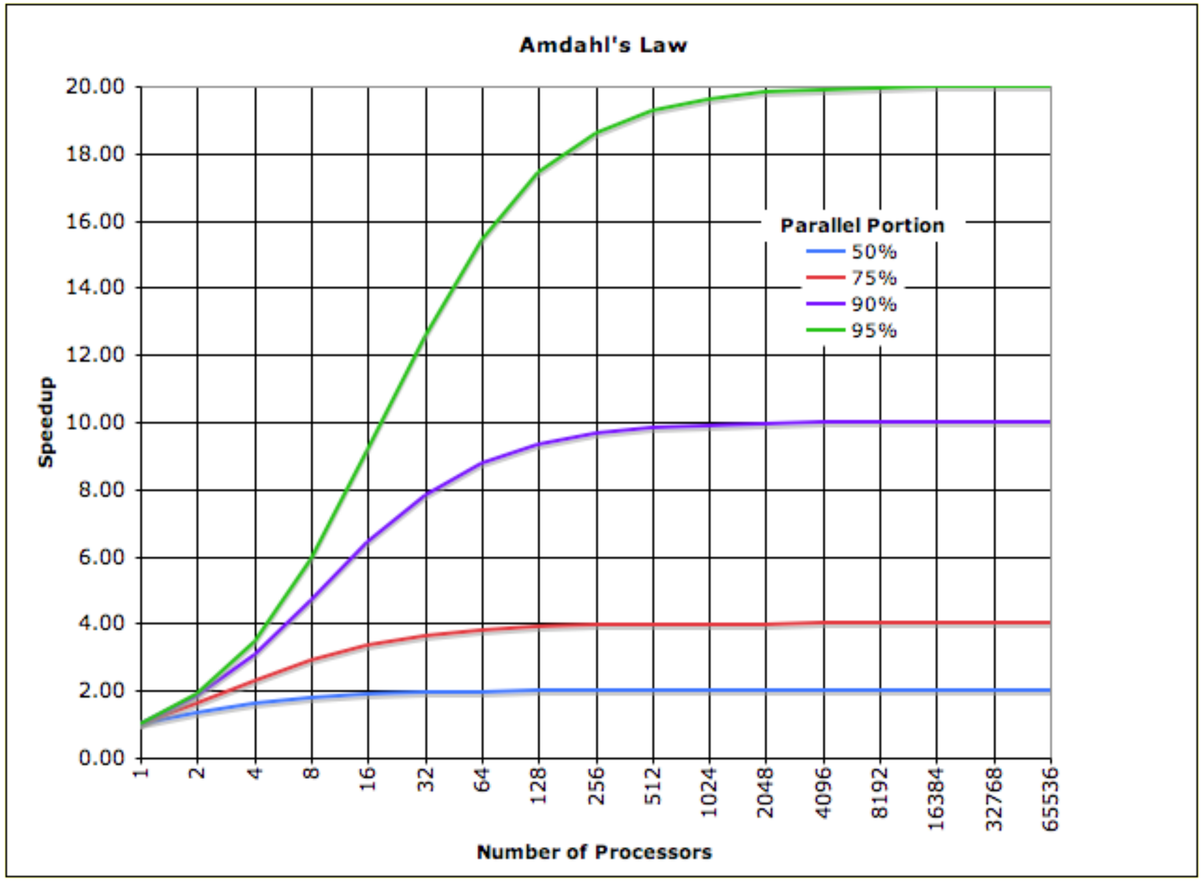
- future: run parts of events in different threads ➜ higher optimization results with even less memory usage



Fermilab

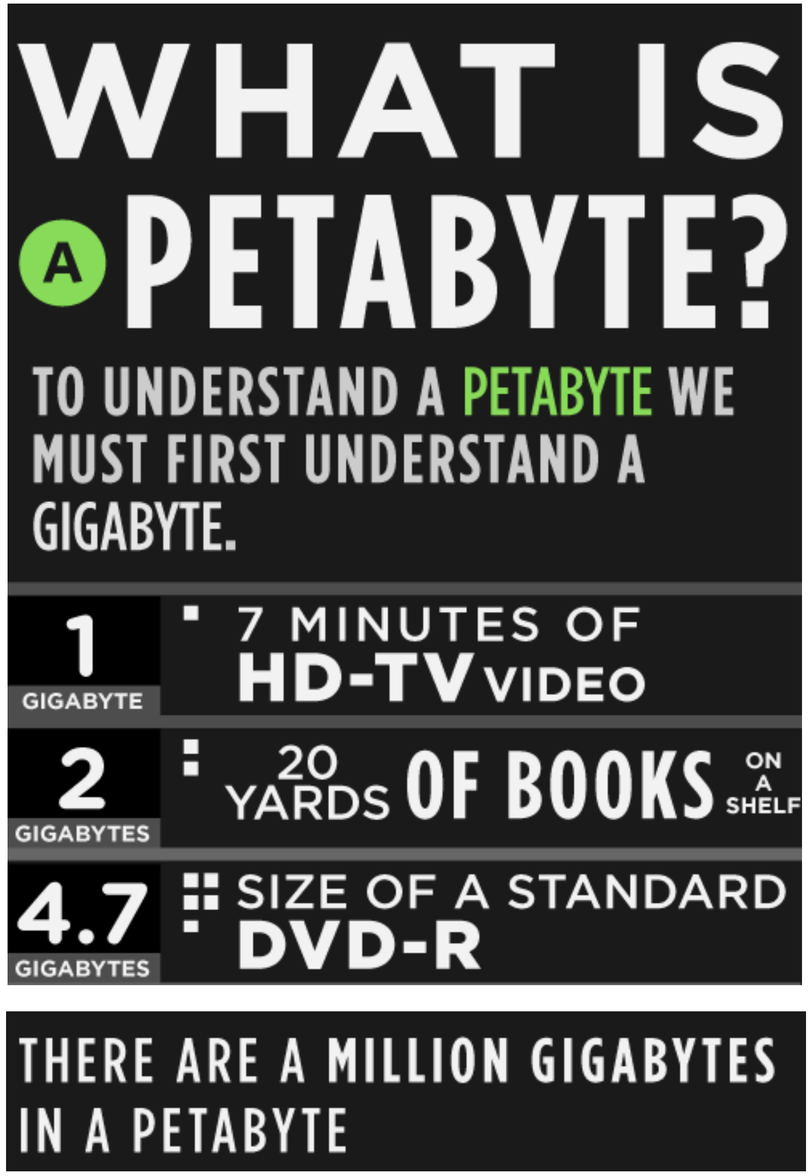
Thread-safe programming

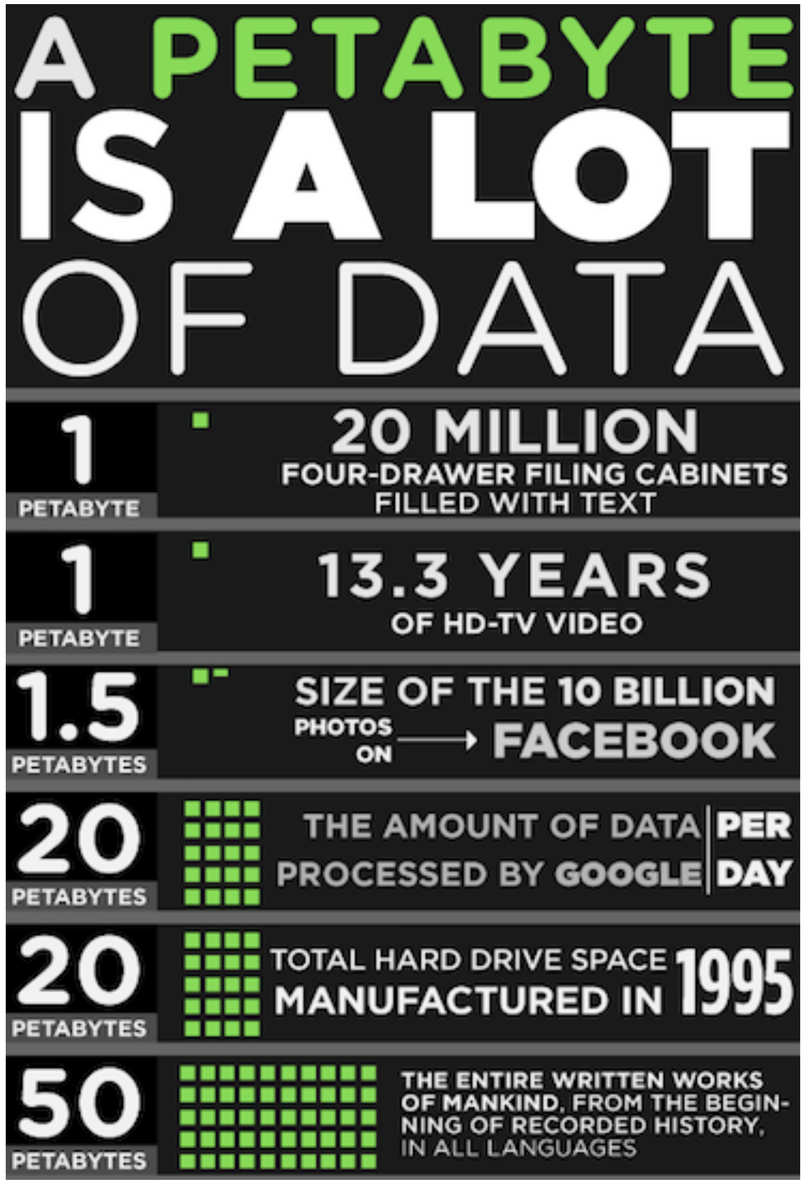
- New technologies: multi-threading, GPGPU, Co-Processors
  - Require new programming skills!
  - My opinion: comparable to Fortran ➜ C++ switch
- Multi-threaded programming needs to be done right
  - Small amounts of non-thread-safe code reduces the efficiency significantly ➜ Amdahl's law
- **Go and learn thread-safe programming!**



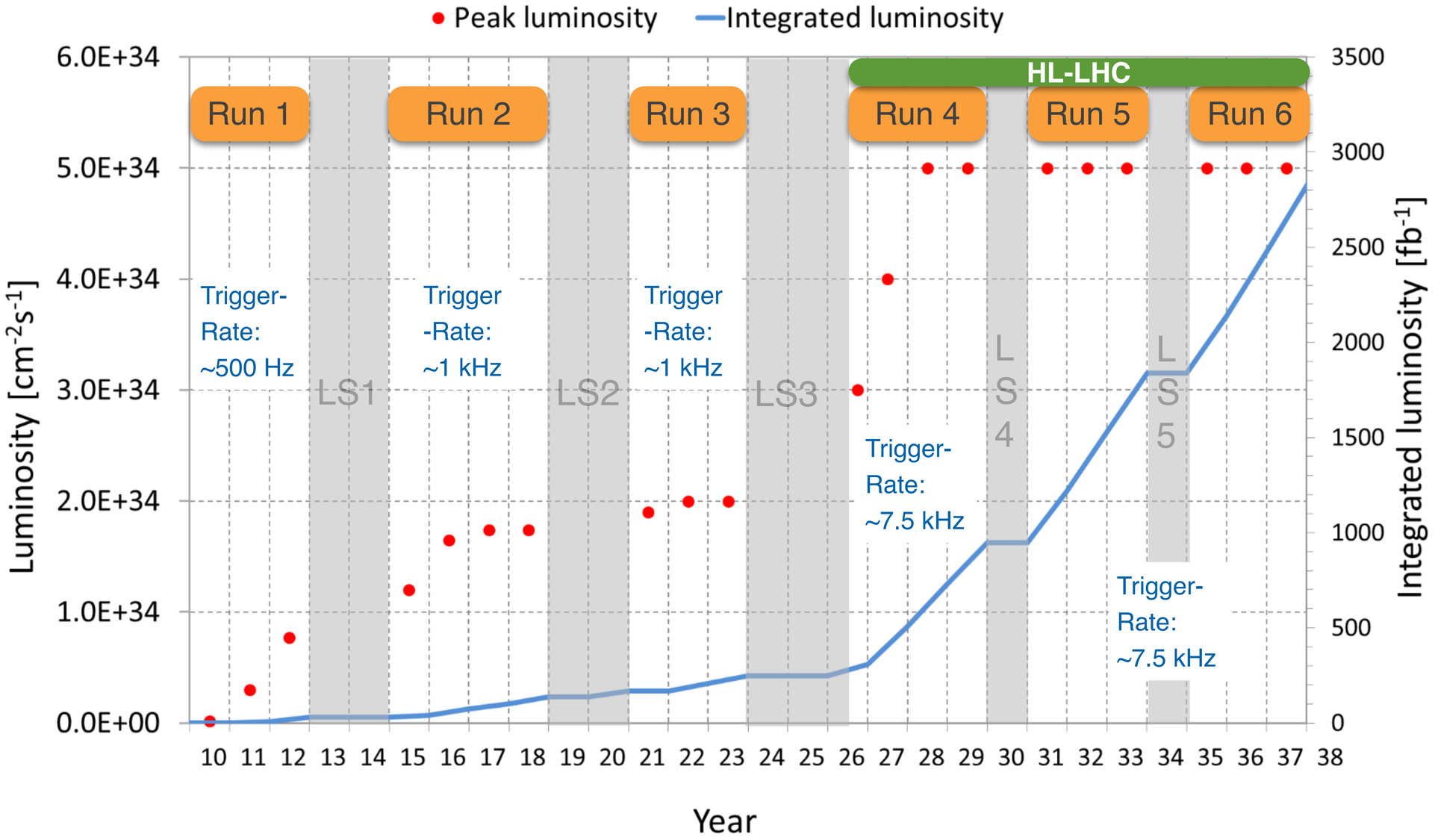
Storage

What is a Petabyte?



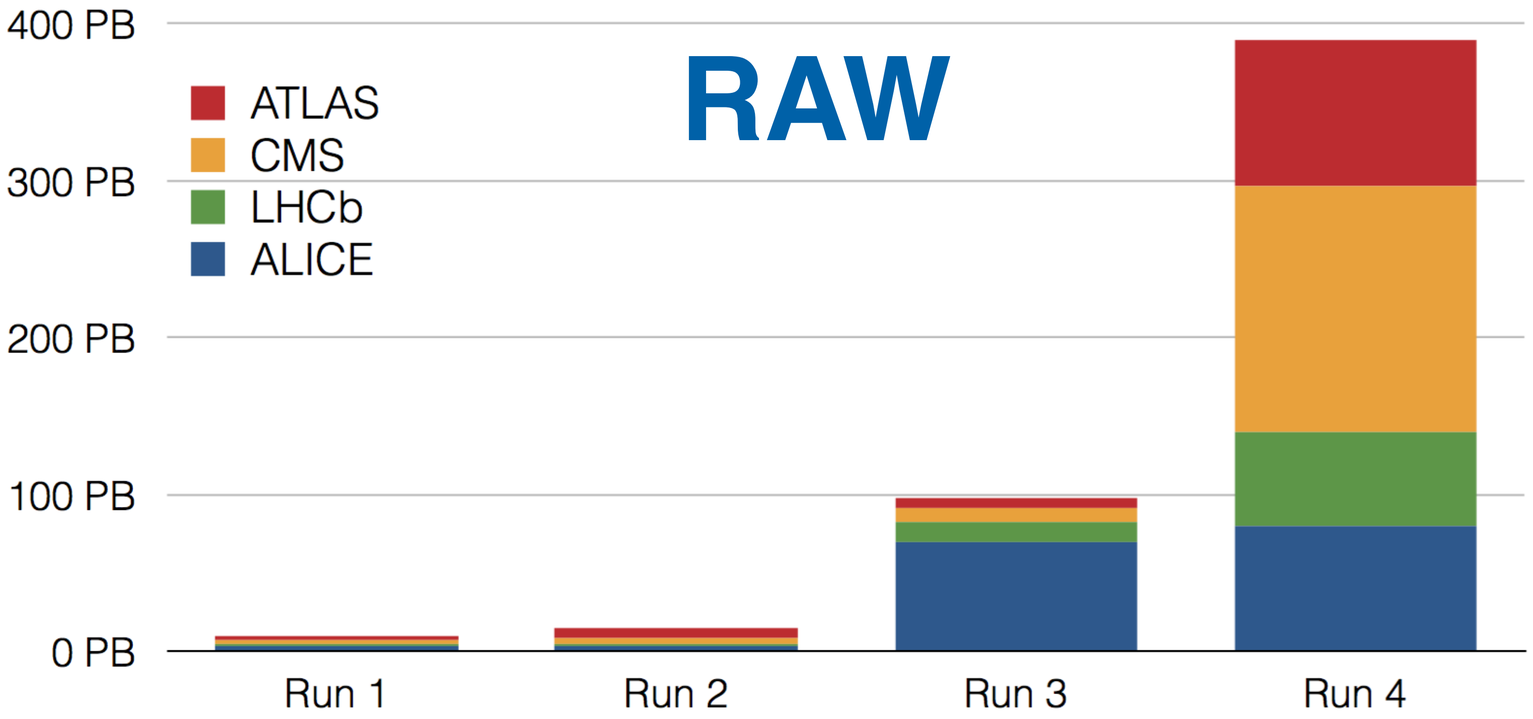


LHC schedule



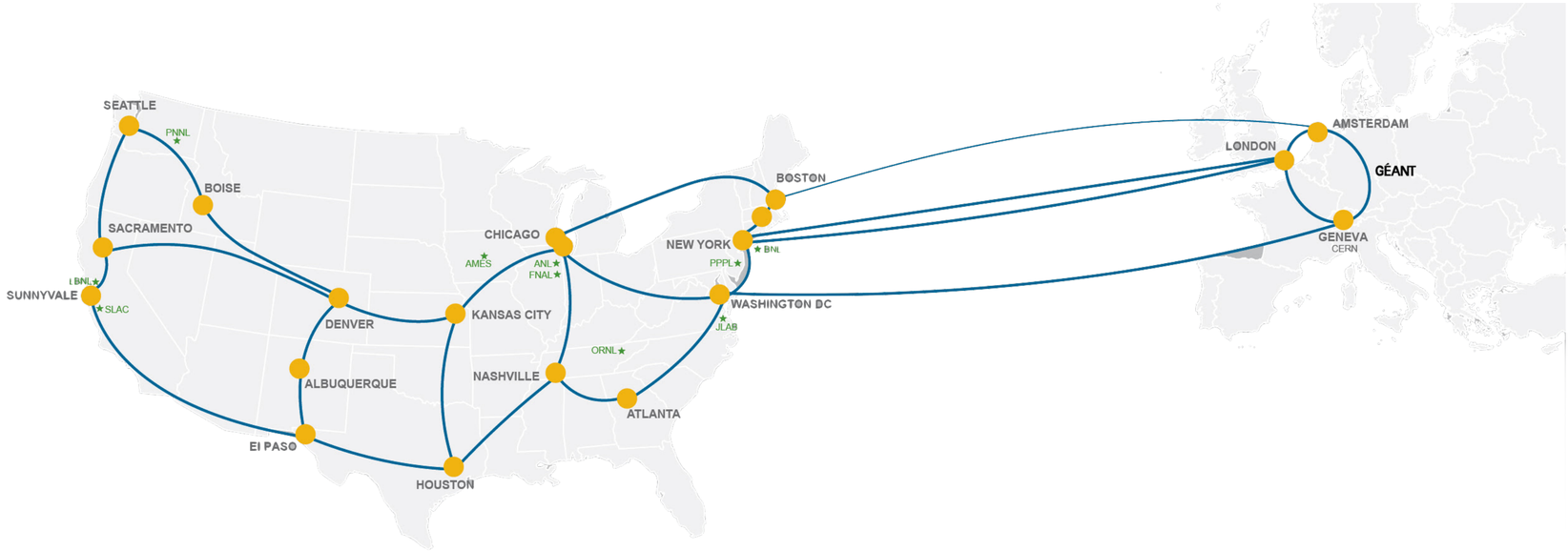
Fermilab

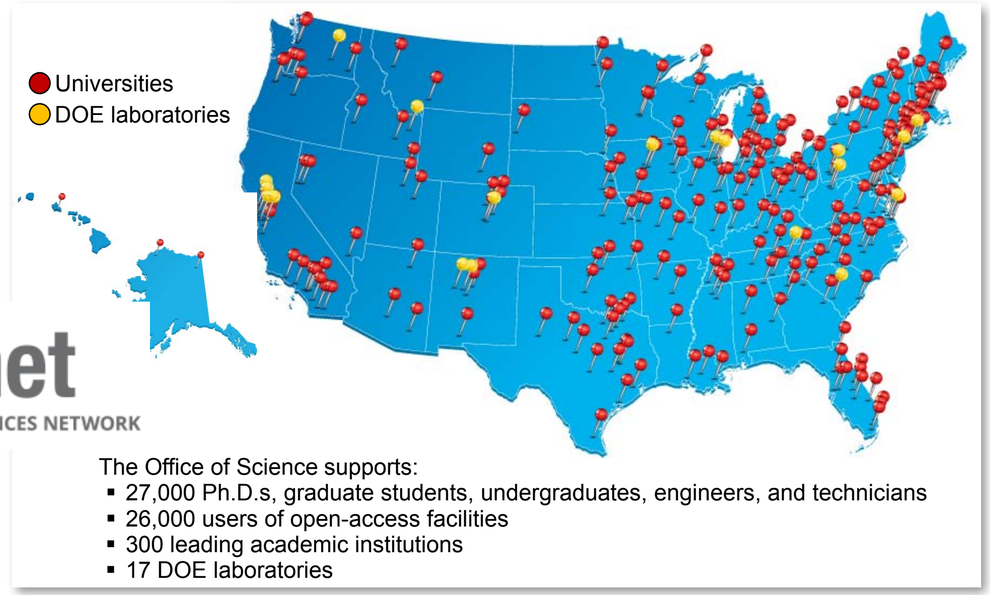
LHC expectation data volumes



- Shown: RAW expectations
  - Derived data (RECO, Simulation): factor 8 of RAW
- LHC Run 4 is starting the exabyte era
- How do we analyze that much data in the future?

Strong networks: ESNet

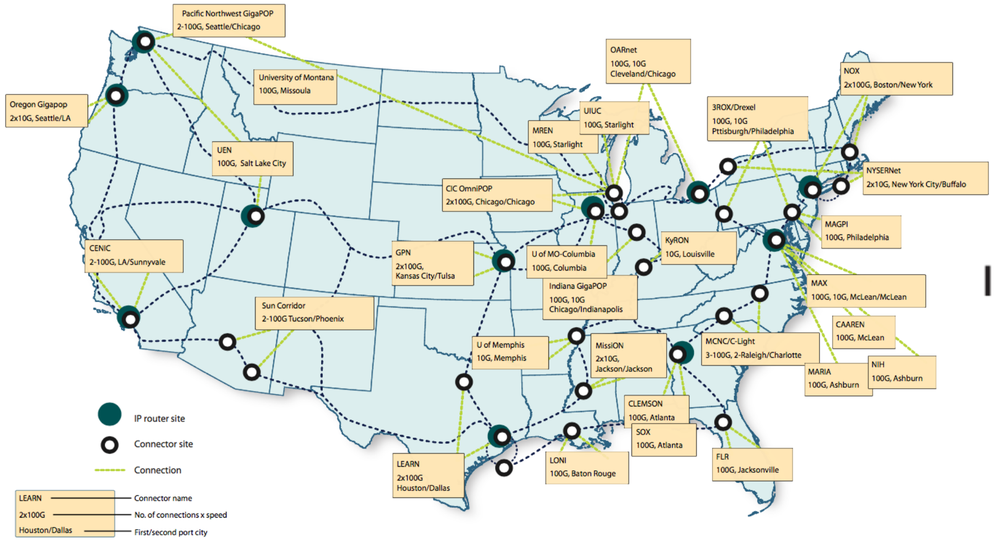




ESnet
ENERGY SCIENCES NETWORK

The Office of Science supports:

- 27,000 Ph.D.s, graduate students, undergraduates, engineers, and technicians
- 26,000 users of open-access facilities
- 300 leading academic institutions
- 17 DOE laboratories

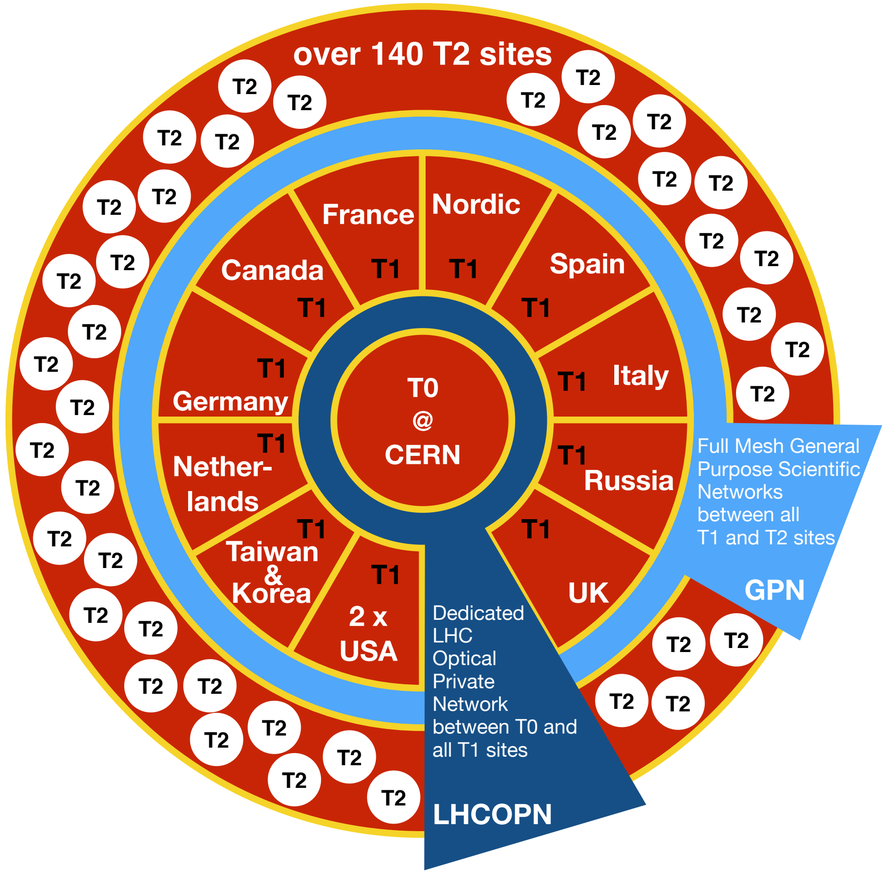


INTERNET 2

Fermilab

Distributed infrastructures and transfer systems

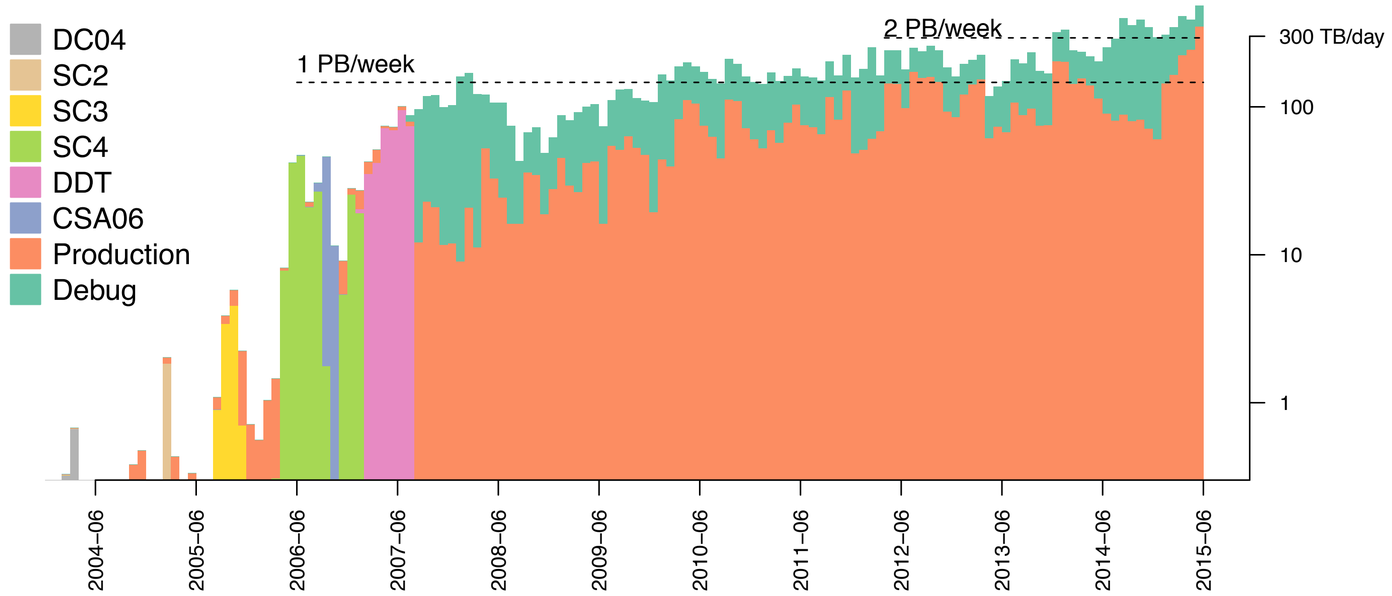
Example: Worldwide LHC Grid (WLCG)



Community uses various solutions to provide distributed access to data:

Experiment specific: Atlas (Rucio), CMS (PhEDEx), …

Shared: SAM (Neutrino and Muon experiments)



CMS transfers: more than 2 PB per week

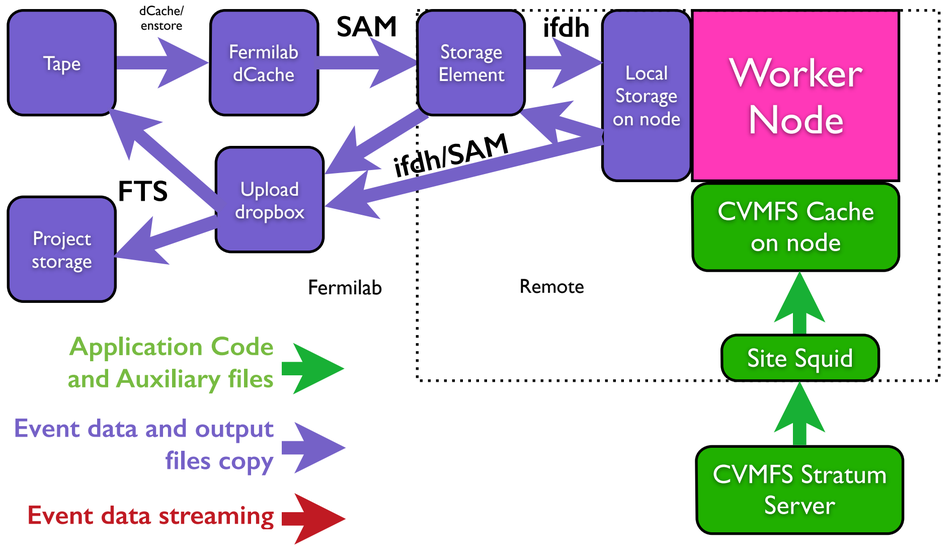
Fermilab

Dynamic Data Management

- Subscription based transfer systems
  - PhEDEx (CMS) and Rucio (Atlas)
  - LHC Run 1: mostly manual operations
  - LHC Run 2: **dynamic data management**
    - Popularity is tracked per dataset
    - Replica count across sites is increased or decreased according to popularity

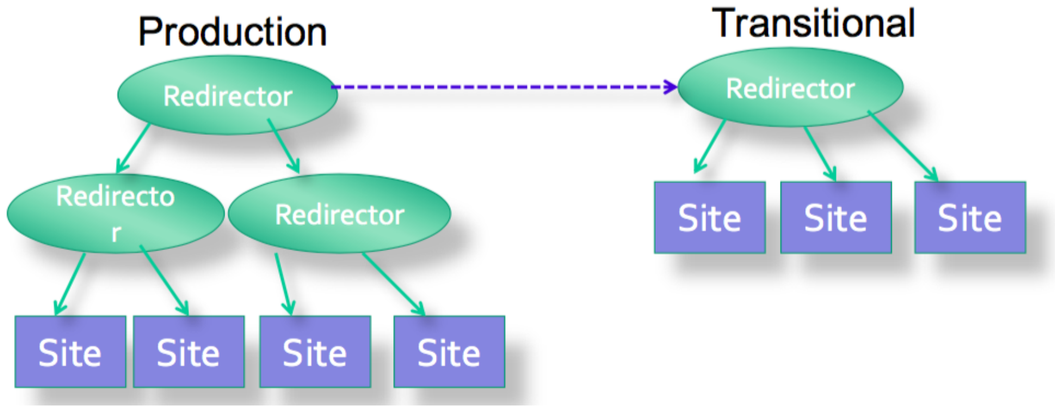
- Fully integrated distribution system
  - SAM (shared amongst Neutrino and Muon experiments)
  - All movement is based on requests for datasets from jobs.
  - Interfaces to storage at sites, performs cache-to-cache copies if necessary

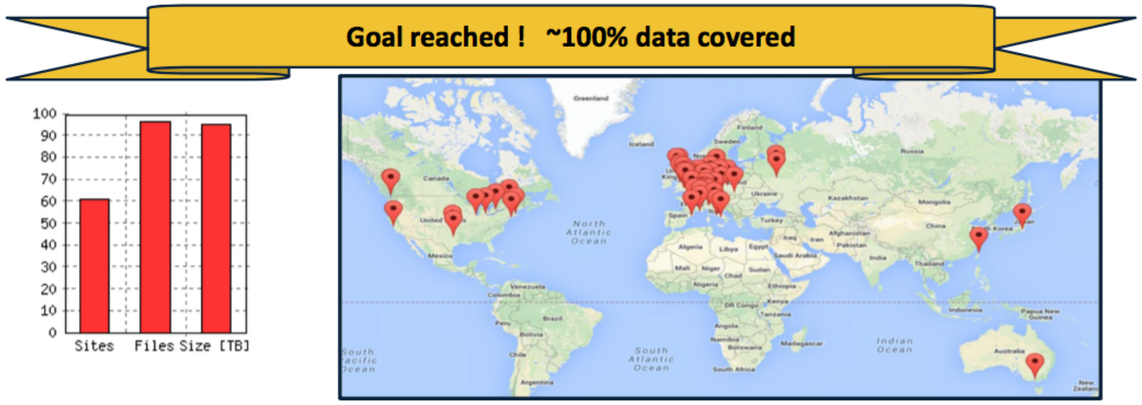
- Data is distributed automatically for the community



Data Federations

- xrootd: remote access to files
- ALICE based on xrootd from the beginning
- CMS and Atlas deployed xrootd federations
  - AAA for CMS, FAX for Atlas
  - Allows for remote access to all files on disk at all sites
  - Use cases:
    - Fall back
    - Overflow for ~10% of all jobs





Fermilab

OSG StashCache

- OSG: StashCache
  - Bringing opportunistic storage usage to all users of OSG
  - OSG collaborators provide local disk space
  - OSG is running xrootd cache servers
    - Dynamic population of caches ➜ efficient distributed access to files
      - For users that don't have infrastructures like CMS and Atlas

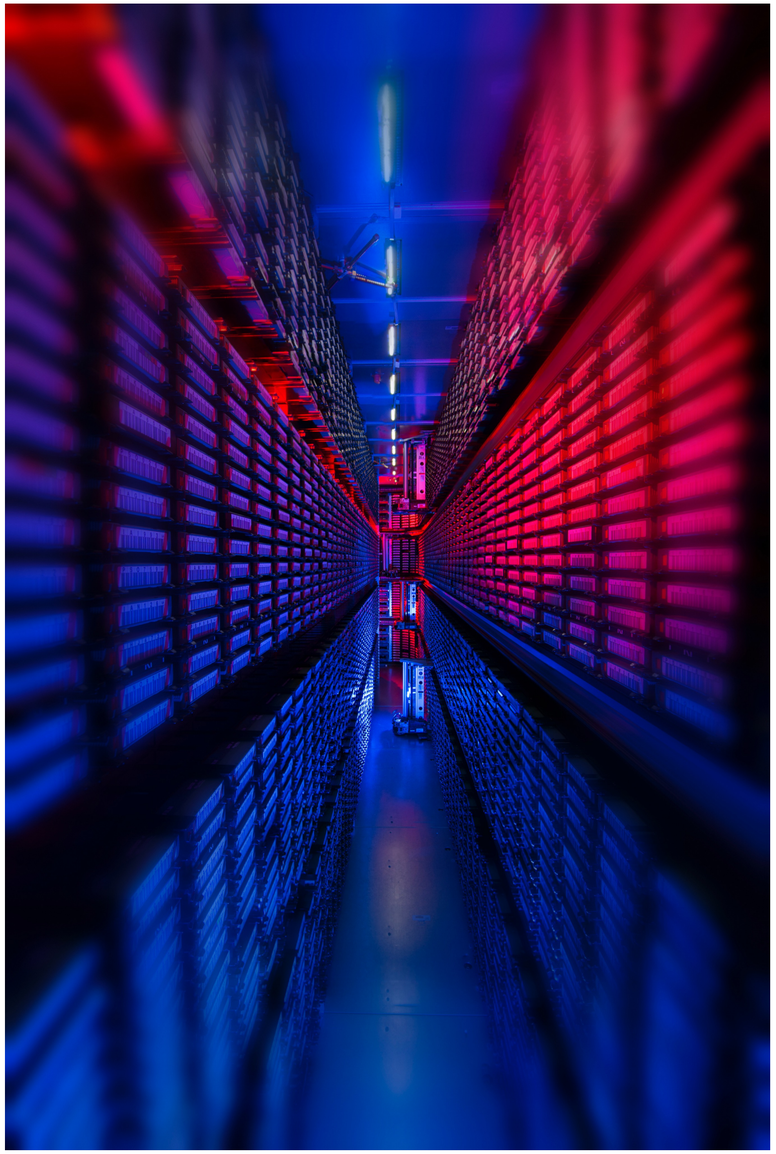
Stash origin: ★

OSG Caches: ●

Fermilab

Active Archival Facility

- HEP has the tools and experience for the **distributed exabyte scale**
  - We are “best in class” in the field of scientific data management
- We are working with and for the whole science community
  - To bring our expertise to everyone’s science
  - To enable everyone to manage, distribute and access their data, globally
- Example: Fermilab’s Active Archival Facility (AAF)
  - Provide services to other science activities to preserve integrity and availability of important and irreplaceable scientific data
  - Projects:
    - Genomic research community is archiving datasets at Fermilab’s AAF and providing access through Fermilab services to ~300 researchers all over the world
    - University of Nebraska and University of Wisconsin are setting up archival efforts with Fermilab’s AAF



Processing

New resource providers

Grid

- Virtual Organizations (VOs) of users trusted by Grid sites
- VOs get allocations ➜ **Pledges**
  - Unused allocations: opportunistic resources

Trust Federation

Cloud

- Community Clouds - Similar trust federation to Grids
- Commercial Clouds - **Pay-As-You-Go** model
  - Strongly accounted
  - Near-infinite capacity ➜ **Elasticity**
  - Spot price market

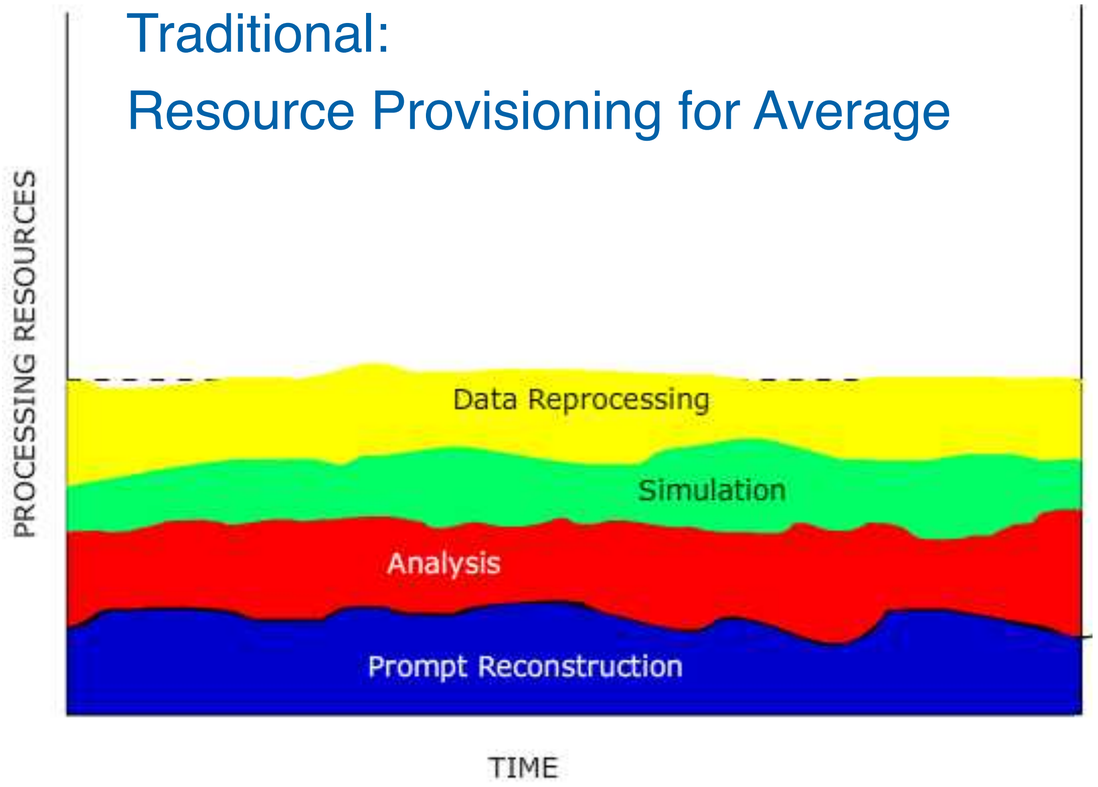
Economic Model

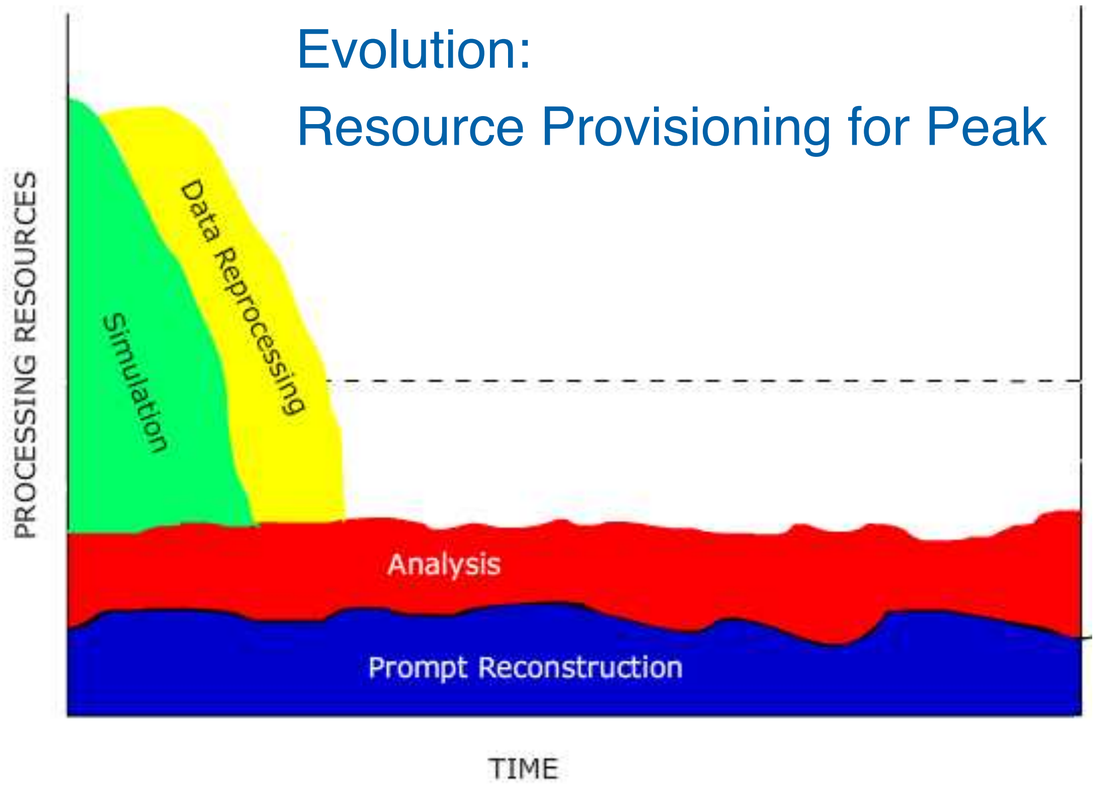
HPC

- Researchers granted access to HPC installations
- Peer review committees award **Allocations**
  - Awards model designed for individual PIs rather than large collaborations

Grant Allocation

Evolving the Grid



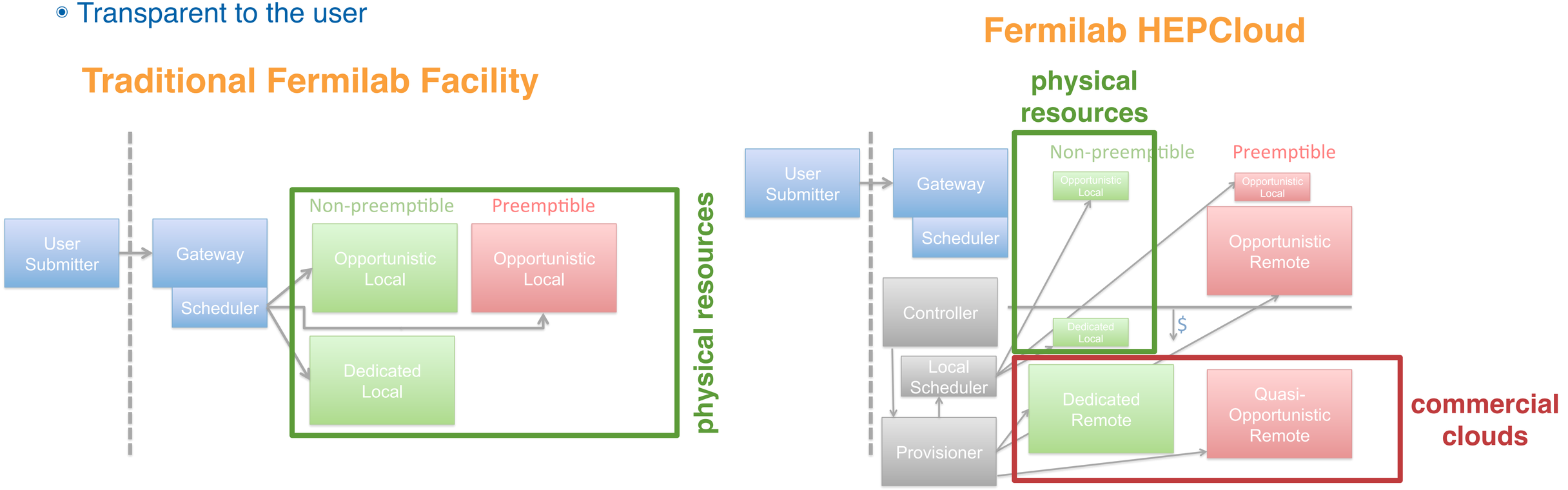


- Experiments don't need all the resources all the time
  - Conference schedule, holiday seasons, accelerator schedules, etc.
  - Resource needs vary with time ➜ Provisioning needs to adapt

Fermilab's HEPCloud

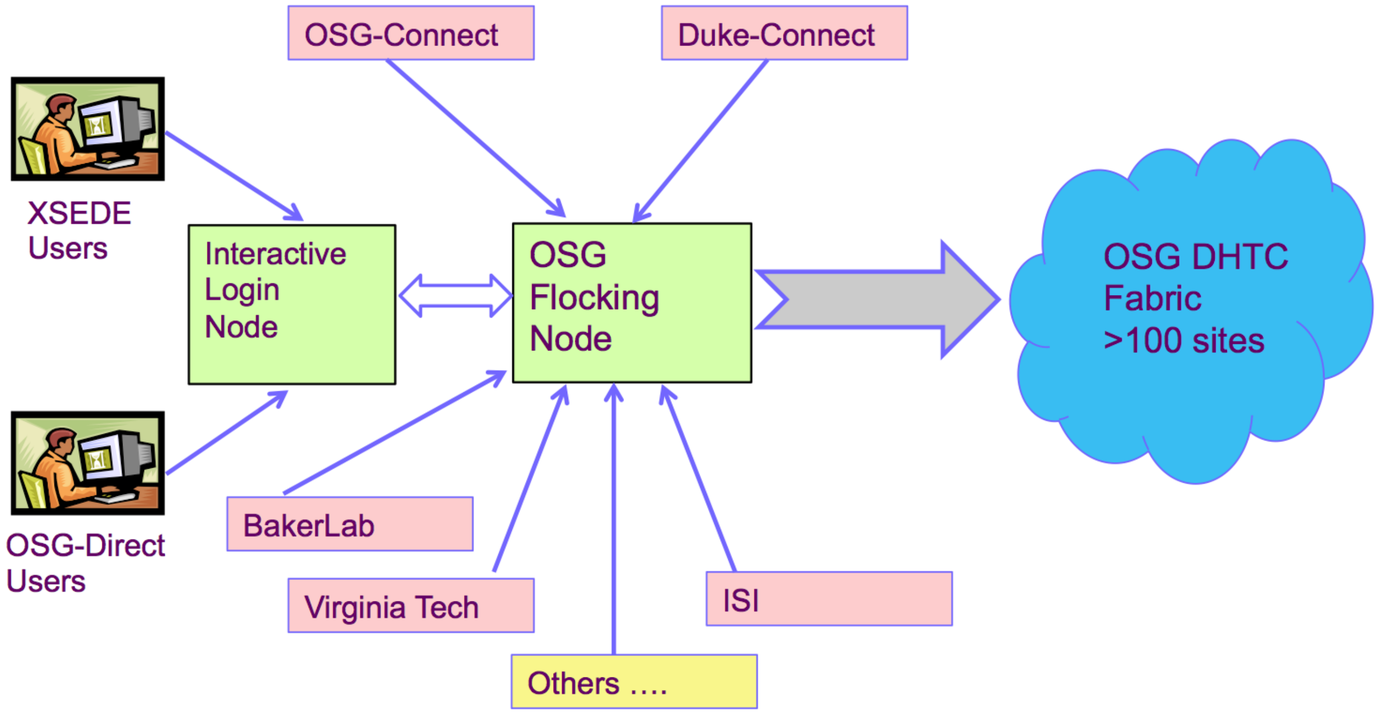
- Many experiments and facilities are exploring using commercial cloud providers to provision for peak
  - Examples: Atlas, CMS, STAR, NOvA, etc. / BNL, FNAL, CNAF, etc.

- Example: Fermilab's HEPCloud
  - Provision commercial cloud resources in addition to physically owned resources
  - Transparent to the user



Open Science Grid ➜ Facilitating shared access

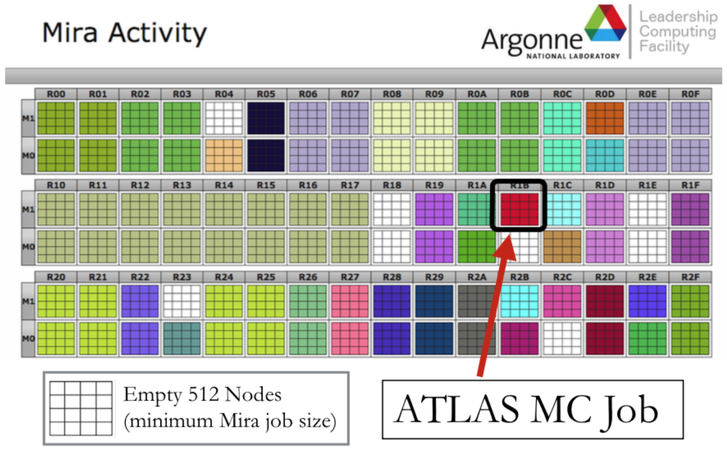
- Researcher use a single interface to use resources …
  - ... they own
  - ... others are willing to share
  - ... they have an allocation on
  - ... they buy from a commercial (cloud) provider



- OSG focuses on making this technically possible for Distributed High Throughput Computing
  - Operate a shared Production Infrastructure ➜ Open Facility (glideinWMS)
  - Advance a shared Software Infrastructure ➜ Open Software Stack
  - Spread knowledge across Researchers, IT professionals & Software developers ➜ Open Ecosystem

HPC & HEP

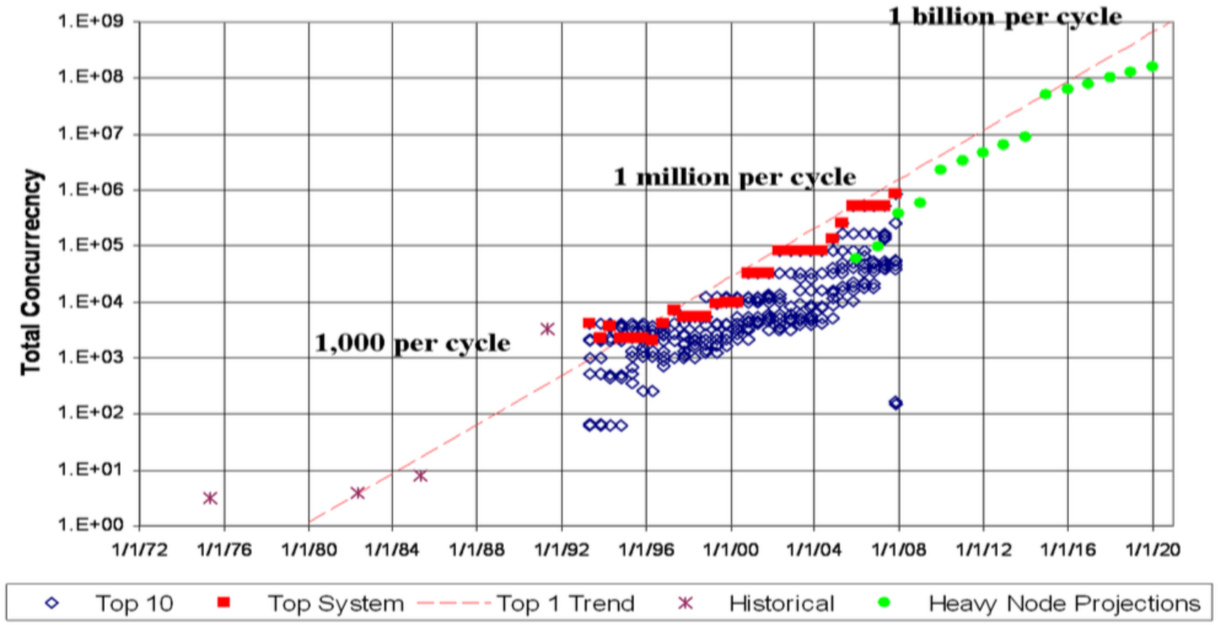
- HTC: High Throughput Computing
  - Independent, sequential jobs that can be individually scheduled on many different computing resources across multiple administrative boundaries(*)
- HPC: High Performance Computing
  - Tightly coupled parallel jobs, must execute within a particular site with low-latency interconnects(*)
- Long history in HEP in using HPC installations
  - Lattice QCD and Accelerator Modeling exploit the low latency interconnects successfully for a long time
- Community effort: enable traditional HEP framework applications to run on HPC installations
  - Example: Mira at Argonne (PowerPC, ~49k nodes each 16 cores, almost 800k cores)
  - Generating Atlas LHC Events with Algren



The Future: Exascale ➜ more cores!

System attributes	NERSC Now	OLCF Now	ALCF Now	NERSC Upgrade	OLCF Upgrade	ALCF Upgrades	
Name Planned Installation	Edison	TITAN	MIRA	Cori 2016	Summit 2017-2018	Theta 2016	Aurora 2018-2019
System peak (PF)	2.6	27	10	> 30	150	>8.5	180
Peak Power (MW)	2	9	4.8	< 3.7	10	1.7	13
Total system memory	357 TB	710TB	768TB	~1 PB DDR4 + High Bandwidth Memory (HBM) +1.5PB persistent memory	> 1.74 PB DDR4 + HBM + 2.8 PB persistent memory	>480 TB DDR4 + High Bandwidth Memory (HBM)	> 7 PB High Bandwidth On-Package Memory Local Memory and Persistent Memory
Node performance (TF)	0.460	1.452	0.204	> 3	> 40	> 3	> 17 times Mira
Node processors	Intel Ivy Bridge	AMD Opteron Nvidia Kepler	64-bit PowerPC A2	Intel Knights Landing many core CPUs Intel Haswell CPU in data partition	Multiple IBM Power9 CPUs & multiple Nvidia Voltas GPUS	Intel Knights Landing Xeon Phi many core CPUs	Knights Hill Xeon Phi many core CPUs
System size (nodes)	5,600 nodes	18,688 nodes	49,152	9,300 nodes 1,900 nodes in data partition	~3,500 nodes	>2,500 nodes	>50,000 nodes
System Interconnect	Aries	Gemini	5D Torus	Aries	Dual Rail EDR-IB	Aries	2nd Generation Intel Omni-Path Architecture
File System	7.6 PB 168 GB/s, Lustre®	32 PB 1 TB/s, Lustre®	26 PB 300 GB/s GPFS™	28 PB 744 GB/s Lustre®	120 PB 1 TB/s GPFS™	10PB, 210 GB/s Lustre initial	150 PB 1 TB/s Lustre®

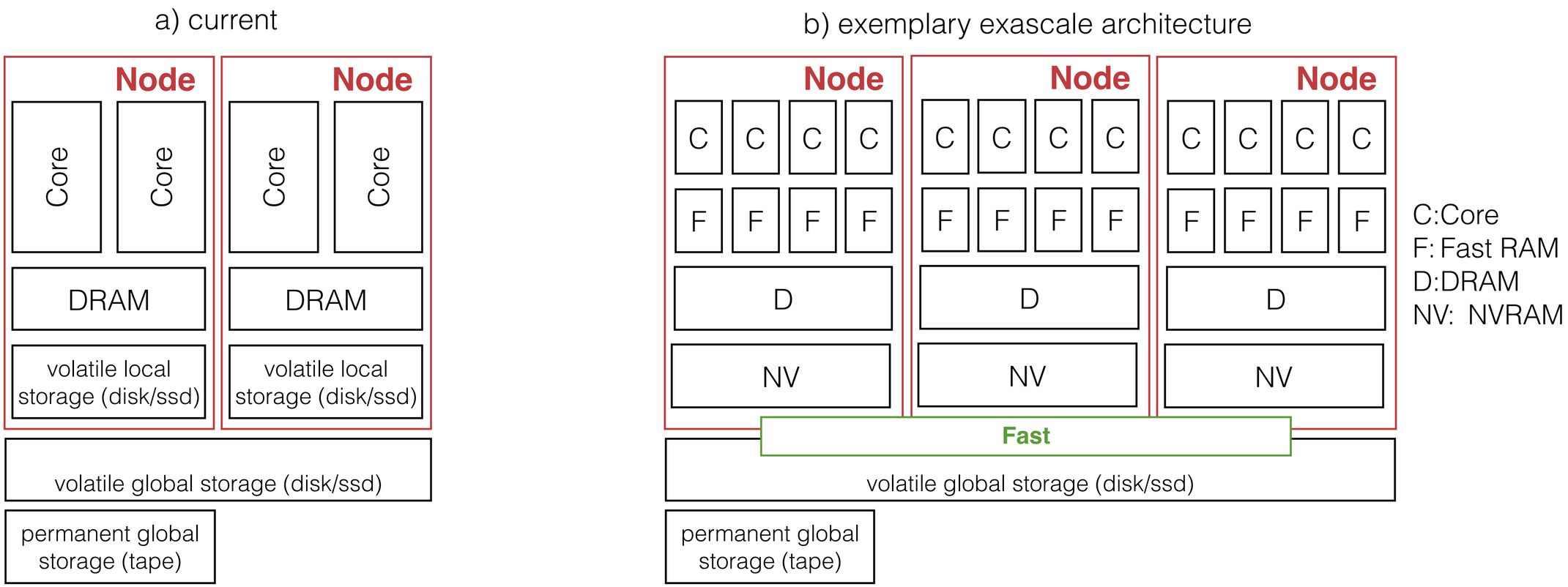
Projected Parallelism for Exascale



- Department of Energy's (DOE) Advanced Scientific Computing Research (ASCR) program plans for Exascale Era ➜ **"A lot more cores!"**
- Opens up exciting possibilities for HEP: in the light of significantly increasing resource needs (for example for the High Luminosity LHC)

Fermilab

New architectures



- HEP applications need a lot of memory and memory bandwidth
  - Cannot have both in Exascale machines ➜ **new architectures**
  - Requires to rethink how we design HEP applications!

Summary & Outlook

Take-home messages

- Software and Computing are integral parts of the HEP science process
  - Know the tools and their capabilities ➜ Get physics results efficiently and reliably
- Learn multi-threaded programming!!!
- Having to handle Exabytes of data is not that far off
  - Many new tools help you, both if you are working for a LHC collaboration, the Neutrino and Muon Experiment Community or any other HEP or non-HEP experiments
- Science will look different in the Exascale era
  - Commercial clouds and Exascale HPC machines will change the way when and how we do computing

Acknowledgements

- Many thanks to DPF 2015 for the invitation.

- Thanks to
  - All my colleagues who make running science software at unprecedented scales possible
  - All my colleagues who helped preparing this talk

And now:

No, lunch